Energy Research and Development Division
FINAL PROJECT REPORT

SEAMICRO VOLUME SERVER POWER REDUCTION RESEARCH AND DEVELOPMENT

Prepared for: California Energy Commission
Prepared by: SeaMicro, Inc.

DECEMBER 2012
CEC-500-2013-003

Prepared by:

Primary Author(s):

Andrew Fetter
Gary Lauterbach
Anil Rao
Mike Partridge

SeaMicro, Inc. |
1237 East Arques Ave.
Sunnyvale, CA 94085
Direct: 408-701-5065 | Fax: 408-701-5097
http://www.seamicro.com/

Contract Number: PIR-10-057

PREFACE

The California Energy Commission Energy Research and Development Division supports public interest energy research and development that will help improve the quality of life in California by bringing environmentally safe, affordable, and reliable energy services and products to the marketplace.

The Energy Research and Development Division conducts public interest research, development, and demonstration (RD&D) projects to benefit California.

The Energy Research and Development Division strives to conduct the most promising public interest energy research by partnering with RD&D entities, including individuals, businesses, utilities, and public or private research institutions.

Energy Research and Development Division funding efforts are focused on the following RD&D program areas:

- Buildings End-Use Energy Efficiency
- Energy Innovations Small Grants
- Energy-Related Environmental Research
- Energy Systems Integration
- Environmentally Preferred Advanced Generation
- Industrial/Agricultural/Water End-Use Energy Efficiency
- Renewable Energy Technologies
- Transportation

SeaMicro Volume Server Power Reduction Research and Development is the final report for the grant number PIR-10-057 conducted by SeaMicro, Inc. The information from this project contributes to Energy Research and Development Division's Industrial/Agricultural/Water End-Use Energy Efficiency Program.

For more information about the Energy Research and Development Division, please visit the Energy Commission's website at www.energy.ca.gov/research/ or contact the Energy Commission at 916-327-1551.

ABSTRACT

Over the past 10 years, the data center has undergone a drastic change in size and scope. Despite these changes, the server remained architecturally unchanged. With the rise of the Internet and the rapid growth of the data center, the challenge is how to handle the huge volume of relatively modest computational workloads. These workloads are generated by the millions of independent users each wanting to search, view web pages, check email, and read the news. Multi-core Central Processing Units (CPU) are particularly inefficient for these simple computational workloads. The mismatch between the CPU in generalist volume servers and specialized workloads in the data center is a fundamental cause of power consumption concerns.

SeaMicro has developed new technologies that include servers the size of a credit card and an innovative supercompute fabric that connects thousands of processor cores, memory, storage, and input/output traffic. These innovative technologies are incorporated into the SeaMicro SM10000 Compute Appliance to replace volume servers in data centers. The SeaMicro SM10000 typically uses one quarter the power and takes one sixth the space of traditional servers with the same compute performance, yet delivers up to 12 times the bandwidth per core.

SeaMicro's innovative technology benefits California by creating jobs. This work also reduces operation costs, enhances performance, and increases energy efficiency in large data centers and computing-intensive environments, resulting in potential savings to California ratepayers.

Keywords: Supercompute fabric, power server, volume server, energy efficient, power reduction, data center, SeaMicro Corporation

Please use the following citation for this report:

Fetter, Andrew, Gary Lauterbach, Anil Rao, Mike Partridge (SeaMicro Corporation). 2012. ***Volume Server Power Reduction Research and Development.*** California Energy Commission. CEC-500-2013-003.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Issue

Electricity consumed by data centers is approximately 3 percent of California's total electrical consumption and is growing rapidly. Energy use by data centers has doubled every five years, and the largest user of energy within data centers is the volume server. Volume servers are the most common server in the data center industry and use about 68 percent of the energy in a data center, excluding support equipment. For companies in the data center, power consumption is the largest operating expense which often accounts for more than 30 percent of operating expenses.

For today's volume servers, the Central Processing Units (CPUs) are not well matched for all workloads in the data center and for the cloud computing environment. Cloud computing is the use of hardware and software computing resources that are delivered as a service over a network. Volume servers use multi-core CPUs that are designed for complex computational challenges and, in terms of computation per dollar, are best-in-class at complex workloads. In comparison, volume servers are not as efficient at simple computational workloads. Volume servers are also called "rack servers" because they are stacked like pizza boxes one on top of another in metal racks. These servers are usually 1.75 inches to 3.5 inches tall, 18 inches across, and up to 30 inches deep. Each server is a discrete unit that is individually managed, controlled, powered, and cooled. These individual servers are linked together with Gigabit Ethernet switches and routers. Gigabit Ethernet is a term describing various technologies for transmitting Ethernet frames at a rate of 1 gigabit, or 1 billion bits, per second. This results in massive duplication of components as each server has the overhead necessary for individual operation, management, and connectivity, but is never used in isolation.

Background

SeaMicro, Inc. (SeaMicro) was awarded a $9.3 million grant in 2010 from the U.S. Department of Energy (DOE) and a $250,000 match Public Interest Energy Research (PIER) grant in 2011 from the California Energy Commission to develop the Volume Server Power Reduction Research and Development Project (Project). The Project has accelerated development and commercialization of a prototype compute appliance to replace volume servers in appropriate situations, which uses significantly less energy and reduces the footprint and cost.

Volume servers reportedly consume more than one percent of the total electricity used in the United States at a cost of more than $3 billion dollars each year. In addition, power consumed by volume servers more than doubled between 2000 to 2006.[1] Power consumption is the largest operating expense for data centers, often accounting for more than 30 percent of operating expenses. Over a volume server's lifetime, the cost of power often exceeds its purchase price.

---

[1] 2007 *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431*.

Over the past 10 years, the data center has undergone a drastic change in size and scope, including dramatic changes in the demand for compute through cloud computing, the type of compute required, and the economics of operation. Many consumers and business services began using cloud computing, and it became a source of on-demand compute. Despite these changes, the server remained architecturally unchanged. Accommodations are needed for the new and different workloads and traffic patterns of the highly specialized workloads that now dominate the data center. The mismatch between specialized workloads and generalist servers is an underlying cause of power consumption concerns in the data center.

SeaMicro saw this opportunity, but unlike previous approaches to reducing power draw in servers, did not focus on redesigning power supplies or improving airflow. Instead, SeaMicro brought together technical insights from CPU design, virtualization, supercomputing, and networking to create a new server architecture optimized for scale-out infrastructures. SeaMicro characterized the specific computational requirements of scale-out work in data centers and designed custom hardware, Application Specific Integrated Circuits (ASICs) and Field Programmable Gate Arrays (FPGAs), and selected CPUs optimized for scale-out workloads. This created the industry's first purpose-built server optimized for highly partitioned workloads such as those found in the web tier, search, and index computation.

SeaMicro was founded by industry veterans who helped build some of the world's largest data centers. The founders have brought together a deep understanding of data center computation including CPU architecture, clustering, networking, and interconnect technologies. SeaMicro is backed by leading venture capitalists including Khosla Ventures, Draper Fisher Jurvetson and Crosslink Capital, as well as leading public corporations. SeaMicro obtained venture capital for match funding totaling $10.7 million.

Issue and Solution

With the rise of the Internet and the rapid growth of the data center, the CPU has few complex problems. The challenge is how to handle the huge volume of relatively modest computational workloads called web-tier workloads. These workloads are generated by the millions of independent users each wanting to search, view web pages, check email, and read the news. Multi-core CPUs are particularly inefficient for these simple computational workloads. The mismatch between the CPU in volume servers and the primary workload in the data center is a fundamental cause of the rapid rise in power consumption by volume servers.
The most efficient CPUs for the most common workload in the data center were not intended for servers at all. They were designed for handheld devices and small laptop computers. These CPUs are simpler designs requiring less power, and more efficient at web-tier workloads. Reportedly they offer more than three times the performance per watt of the large multi-core CPUs for this workload, and provide half the single-thread performance for a sixth of the power draw, a dramatic improvement in computation per unit of power.

By identifying a more efficient CPU for the workload of a data center, SeaMicro addressed an inherent inefficiency of server packaging used in simple computation. SeaMicro invented input/output virtualization technology that allows common components to be shared by hundreds of CPUs. This allows SeaMicro to drive down the number of the components of the server, everything except the CPU and memory, reducing the server to the size of a credit card.

Seizing upon market demands and its innovative intellectual property, SeaMicro developed an interconnected fabric to link 512 servers in a three-dimensional array inside the compute appliance. SeaMicro's supercompute fabric connects thousands of processor cores, memory, storage, and input/output traffic. The company's fabric supports multiple processor instruction sets. The CPUs are managed with load-balancing software for dynamic allocation of workloads. The intelligent allocation of work to CPUs, based on power-use metrics, allows the CPUs to operate in a vastly more efficient range, while allowing CPUs not being utilized to enter deep sleep.

Results

SeaMicro developed a prototype for an innovative compute appliance that is currently in both production data centers and beta testing with customers. These innovative technologies are incorporated in the SeaMicro SM10000 Compute Appliance to replace volume servers in data centers. The SeaMicro SM10000 typically uses one quarter the power and takes one sixth the space of traditional servers with the same compute performance, yet delivers up to 12 times the bandwidth per core. SeaMicro servers are plug and play—they require no changes to software operating systems, applications, or management infrastructure. The systems are also less expensive to manufacture.

SeaMicro was awarded 2 patents for its work and led the way for industry leaders to reconsider the market for low power servers and create new product lines. The market has been reshaped by the company's unique approach and ground breaking technology.

The PIER match funding from the California Energy Commission, coupled with the DOE grant, was used to:

- Accelerate development engineering and quality assurance testing.
- Enhance dynamic power-management software to further drive down power draw.
- Accelerate product availability.
- Accelerate market adoption by reducing cost and modifying the system to address markets beyond volume servers, such as volume servers working on less-common workloads and mid-range and high-end servers.

The compute appliance addresses the following issues concerning the inefficiency of the CPU within volume servers:

- Different and varying workload requirements by the CPU due to the internet.
- Inefficiency in server packaging for simpler compute tasks that leave enormous duplication in components.

Using the United States Environmental Protection Agency's (U.S. EPA's) estimated energy use for data centers of 61 billion kilowatt-hours/year in 2006, the estimated energy savings in

California would be 3.7 billion kilowatt-hours/year, based on 80 percent of volume servers used for web-tier workloads, 68 percent of energy in the data center used for powering and cooling data centers, and 15 percent of national data center floor space located in California.

Benefits to California

SeaMicro's technologies have reduced costs of operation, and increased performance and energy efficiency in large data centers and cloud computing environments. Mozilla and eHarmony are two California customers successfully using SeaMicro's technology. Numerous other non-public California customers have successfully used the SeaMicro product in test and production facilities, providing multiple benefits to California ratepayers through enhanced energy efficiency. More than 50 California jobs were created at SeaMicro as a result of this project, including hardware, software, manufacturing engineering positions, sales, and sales engineering jobs. The positions have allowed SeaMicro to significantly accelerate engineering development and accelerate commercialization. As a benefit to California's economy, commercialization and delivery to market are months ahead of initial schedule.

CHAPTER 1: SeaMicro Technology Overview

Data centers consume 3 percent of the total electricity in the California. In the data center, 68 percent of the power consumed by Information Technology infrastructure is consumed by volume servers. Volume servers use seven times as much power as the second leading consumer of power: data-networking equipment. According to the 2007 *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431*, between 2000 and 2006, the power used by volume servers in the nation's data centers more than doubled.

SeaMicro developed a revolutionary server: current systems featuring SeaMicro technology typically use one quarter the power and take one sixth the space of traditional servers with the same compute performance, yet deliver up to 12 times the bandwidth per core.[2]

SeaMicro reconceived the volume server as an ultra high density low power single box cluster, a "rack in a box." The system is built around a parallel array of 512 independent ultra low power Intel Atom processors and has a programmable traffic controller front end. The system integrates into a single box—compute, switching, server management, and load balancing.

Three primary technical innovations are key:

- SeaMicro invented and patented technology called CPU input/output, (CPU I/O), virtualization, which dramatically reduces the power draw of the non-CPU portion of a server by driving down the number of components from the motherboard. CPU I/O virtualization allows SeaMicro to shrink the motherboard to the size of a credit card; thereby enabling hundreds of more power efficient CPUs to replace traditional power hungry multi-core processors.

- SeaMicro designed a super computer style interconnected fabric that can link 512 of these credit-card-sized motherboards into a single system with an order-of-magnitude reduction in power draw. SeaMicro developed dynamic routing algorithms to move traffic across the fabric to avoid congestion and to route around failure. This fabric provides an order of magnitude reduction in power over today's communication technologies while providing lower latency, lower cost and more bandwidth. The fabric enables the SeaMicro architecture to support any CPU instruction set and protocol—Ethernet, fibre channel, data center Ethernet, etc.

---

[2] ¼ the power based on a comparison of same throughput achieved by 51 traditional 2P 1U rack servers @ 13,581 total watts at 100% utilization and 512 1P SeaMicro servers in a single 10U chassis at 2,499 total watts at 100% utilization, running a standard web server workload. 1/6 the floor space is based on a comparison of 60 1U traditional servers versus one SeaMicro chassis at 10U. 12X throughput is based on a traditional Dual socket platform with 12 cores (2 socket x six cores) and 2x1GB NICs (2 Gig/12 cores = 167 Mbps bandwidth per core) compared to a SeaMicro single socket server 4 cores and aggregated bandwidth of up to 8 1 Gig NICs for each socket (8 Gig/4 cores = 2Gbps bandwidth per core) 2/.167 = 12, NIC = Network Interface Card

• SeaMicro invented technology that combines CPU management and load balancing, allowing the company to dynamically allocate workloads to specific CPUs on the basis of power-usage metrics. This achievement ensures that the active CPUs operate in a vastly more energy efficient range of utilization. In addition, it allows the user to create pools of CPUs for a given application and can then dynamically add compute resources to the pool based on predefined utilization thresholds.

1.1 Background

The power consumed by servers is a major issue at the micro and macro economic levels. For the data center owner, power as it is often the single largest single operating expense line item accounting for more than 30 percent. At the national level, the problem is just as severe as detailed in the 2007 *EPA Report to Congress on Server and Data Center Efficiency, Public Law 109-431.*

The energy used by the nation's servers and data centers is significant. It is estimated that this sector consumed about 61 billion kilowatt-hours (kWh) in 2006 (1.5 percent of total U.S. electricity consumption) for a total electricity cost of about $4.5 billion.[3]

One type of server—the volume server—was responsible for the majority (68 percent) of the electricity consumed by IT equipment in data centers in 2006. The energy used by this type of server more than doubled from 2000 to 2006, which was the largest increase among different types of servers. The power and cooling infrastructure that supports IT equipment in data centers also uses significant energy as well, accounting for 50 percent of the total consumption of data centers. Among the different types of data centers, more than one-third (38 percent) of electricity use is attributable to the nation's largest (i.e., enterprise-class) and most rapidly growing data centers.[4]

Designed to replace volume servers, the SeaMicro compute appliance delivers significant advantages in power, size, and total cost of ownership:

• Power efficiency: Current systems featuring the SeaMicro compute appliance typically use one quarter the power;
• Size: Current systems featuring the SeaMicro compute appliance take one sixth the space of traditional servers with the same compute performance, yet deliver up to 12 times the bandwidth per core; and
• Total Cost of Ownership: SeaMicro believes that its compute appliance is less costly to buy, own, operate, and manage, than any server currently on the market.[5]

---

[3] *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431, U.S. Environmental Protection Agency ENERGY STAR Program , August 2, 2007*
[4] Id.

[5] ¼ the power based on a comparison of same throughput achieved by 51 traditional 2P 1U rack servers @ 13,581 total watts at 100% utilization and 512 1P SeaMicro servers in a single 10U chassis at 2,499 total watts at 100% utilization, running a standard web server workload. 1/6 the floor space is based on a comparison of 60 1U traditional servers versus one SeaMicro chassis at 10U. 12X throughput is based on a traditional Dual socket platform with 12 cores (2 socket x six cores) and 2x1GB NICs (2 Gig/12 cores = 167

To begin, SeaMicro made several key observations about changes in the computational landscape and weaknesses in the existing architecture of volume servers.

First, SeaMicro recognized that in the data center, and in particular for volume servers in the data enter, the computational requirements had significantly diverged from other parts of the computational landscape. Historically, the challenge in computing was increasingly taxing workloads and led the makers of CPUs to focus on “single-thread performance”— the ability to speed up work on a single hard problem. With the rise of the Internet and the rapid growth of the data center and Cloud Computing, this has changed. Rather than relatively few complex problems, the computational challenge, in the part of the data center dominated by volume servers, became how to handle the huge volume of relatively modest computational workloads. These workloads are generated by the millions of independent users each wanting to search, view web pages, check email, and read the news, all for free.

With companies such as Yahoo!, Google, Facebook, Wal-Mart, New York Times, and others growing to hundreds of millions of users, the challenge for the volume server in their primary application became “how to handle small computational workloads at a scale that has never before been encountered?” While this challenge appeared first in Internet sites, it now has come to dominate corporate data centers, as large enterprises increasingly provide services to their customers and employees over the Internet.

The exponential growth of this workload produces a challenge for today’s volume servers: their CPUs are not well matched for the workload that has come to dominate the data center. Volume servers have historically used large, complex, high speed multi-core CPUs. These CPUs are designed for complex computational challenges and, in terms of computation per dollar, are best-in-class at complex workloads. But these same large, complex, high speed multi-core CPUs are particularly inefficient at small simple computational workloads. Large, complex, high speed multi-core CPUs are “overkill” in these situations. Indeed, the mismatch between the CPU in volume servers and the primary workload in the data center is a fundamental underlying cause of the rapid rise in power consumption by volume servers.

A second driving force underpinning the power issue in the data center is the packaging of the volume servers themselves. Volume servers are also called “rack servers” because they are stacked like pizza boxes one on top of another in metal racks. These servers are usually 1.75”–3.5” tall, 19” across, and up to 30” deep. Each server is a discrete unit—it is individually managed, controlled, powered, and cooled. These individual servers are linked together with Gigabit Ethernet switches and routers. Herein lies the second source of the power issue in the data center—massive replication of components—as each server has the overhead necessary for individual operation, management, and connectivity but is never used in isolation. It is as if a train had been constructed with each train car having its own engine and its own caboose and then linked together to form the train. Obviously, Americans do not move freight that way.

Mbps bandwidth per core) compared to a SeaMicro single socket server 4 cores and aggregated bandwidth of up to 8 1 Gig NICs for each socket (8 Gig/4 cores = 2Gbps bandwidth per core) 2/.167 = 12

Rather we use one engine and one caboose, and amortize these "overhead costs" over hundreds of train cars.

To summarize, the power issue in the data center was born of the confluence of several factors:

(a) The rise to dominance of a new and different workload and Cloud Computing;

(b) The fact that the CPUs that had historically been the data-center workhorse were suddenly ill-suited for the predominant and fastest-growing workload;

(c) The inherent inefficiency in server packaging for simple computation that left enormous duplication in components.

The SeaMicro compute appliance systematically addresses each of these factors to better ensure that the computation delivered is in complete alignment with the most prevalent workload. To execute on this architecture, SeaMicro innovations span multiple domains: communication networks, supercomputer design, routing, ASIC design, and motherboard design, to name a few.

1.2 Technology

The observations previously described laid the groundwork for the SeaMicro solution. These observations also shed light on the breadth of the innovations required in order to successfully research, develop, and bring to market this solution.

First, SeaMicro identified a more efficient CPU, one that was better aligned to handle the most common workloads in the data center. SeaMicro then developed technologies that enable these new low-power CPUs to replace the large, complex, high speed multicore CPUs in volume-server applications. This meant rethinking the architecture of the server and reconstituting it as a delivery system for low-power CPU cycles.

Second, SeaMicro developed technology that links hundreds of these CPUs together so that the overhead costs of management and connectivity were amortized over hundreds of CPUs rather than two or four as had been done traditionally—thereby eliminating the replication of components inherent in clusters of volume servers.

Third, SeaMicro developed software and integrated circuit technologies that would allocate load dynamically across its array of parallel-processing CPUs, in order to minimize power usage by ensuring that the CPUs were operating in their maximally efficient range (CPUs, like many electrical systems, are particularly inefficient at low utilization).

For the first challenge, identifying the most efficient CPUs in terms of computation per watt for the most common workload in the data center, it turns out that the most efficient CPUs were not intended for servers at all. Rather, they were designed for handheld devices and the smallest of laptop computers; they are simpler designs, they use less power, and they are significantly more power-efficient at web-tier workloads. In fact, these CPUs offer more than three times the performance per watt of the large multi-core CPUs for this workload. To be more specific, they provide half the single-thread performance for a sixth of the power draw, a dramatic improvement in computation per unit of power. It is important to note that these

CPUs are smaller and slower and do less work per CPU in absolute terms, but for the less demanding computation needs that dominate the Internet data center, they offer fundamental improvements in computation per unit of power. Although smaller, simpler, and slower often means better performance per unit of energy, it also presented substantial engineering challenges at the system level.

For example, when these low-power CPUs are placed into the existing volume-server architecture, the power consumed to do a unit of work actually increases. The CPU in a server uses approximately a third of the total power consumed by the server. As a thought experiment, assume a magical CPU that delivers half the performance of a traditional CPU but uses no power at all. If that CPU were placed into an existing architecture, then the new server would offer half the performance, for two-thirds the power. In other words, it would deliver lower performance per watt than did the original server. In some sense, this is not surprising. In the 70's and early 80's, car makers did not take smaller, more-efficient motors and put them in the same cars that had historically struggled with gas mileage. Rather, they had to rethink the rest of the car as well, and make it smaller, lighter, etc., in order to be well matched to the more efficient motor.

It is the interaction between power reduction in the CPU, and the power reduction in the rest of the system that produces the dramatic gains in computation per watt that SeaMicro has been able to achieve. The key observation is that if one wants to use these small, more-efficient CPUs, then technology must be developed that reduces the non-CPU "two-thirds of the total power" drawn by the server. That is, one needs to scale the reduction in power draw from the CPU across all components. SeaMicro accomplishes this through its technological breakthrough called hardware-based CPU I/O virtualization.

1.3 Hardware-based CPU I/O Virtualization Technology

SeaMicro has developed I/O (input/output) virtualization technology that removes all non-CPU/memory components from the motherboard, while allowing the CPUs to run standard operating systems and software without requiring modification or recompilation. The CPU I/O virtualization technology also allows common components to be shared across hundreds of CPUs, rather than being duplicated on each motherboard. Thus, components such as basic input/output system (BIOS), external network access, storage, and console are instantiated once and then are amortized over the entire system. Hardware-based CPU I/O virtualization enables SeaMicro to eliminate 90 percent of the components from the server and to shrink the server to the size of a credit card.

1.4 Supercomputer-style Interconnected Fabric

Once the server had been shrunk to the size of a credit card, technology needed to be developed that could link together hundreds of these card-sized computational nodes, with significant reductions in power, cost, and latency. For answers, SeaMicro looked to the techniques used to interconnect the CPUs of the largest and most complicated supercomputers and set about scaling the technology down for data-center applications. The result is the SeaMicro interconnect fabric, which ties together 512 computational nodes. It is a three-dimensional torus, with both path redundancy and diversity. The fabric is FLIT-based and wormhole-routed, with integrated virtual-channel technology to manage congestion, and has a throughput of 1.2 Terabits. These technologies combine to produce a low-latency, high-bandwidth, redundant fabric at very low cost. While the fabric has its origins in the supercomputer world, SeaMicro tuned the design of the fabric, and optimized it for the requirements of the data center.

1.5 Dynamic Compute Allocation Technology™ (DCAT)

DCAT adds a layer of intelligence to the distribution of work to the 512 CPUs by combining CPU management and stateful load balancing. The SeaMicro management software is aware of the health of, and the workload on, all the CPUs in the system. Based on this knowledge, the management software can transparently provision the CPUs and dynamically program the field programmable gate arrays (FPGAs) to direct traffic to one group of CPUs and away from another. The technology works by creating Virtual IP addresses which can be assigned to pools of compute as small as one server and as large as 512 servers. The stateful hardware load balancer then distributes flows across these pools of servers using various load-management algorithms including round-robin, least connections, and max connections to enable the internal servers to be used effectively at the most optimal power consumption levels without degradation of end user performance. The pools of compute are accessed as if they were a single CPU. CPUs can be added or removed from a Virtual IP pool dynamically based on predetermined rules. So for example, traffic can be directed to a pool of CPUs to ensure they are operating in the maximally efficient range, while allowing other CPUs to enter deep sleep mode or even to be turned off. Similarly, a utilization threshold for a pool of compute can be set, and if met, CPUs can be dynamically provisioned and added or removed from the pool.

1.6 ASIC and Custom Silicon

SeaMicro instantiates its fabric and CPU I/O virtualization technology in custom application-specific integrated circuits (ASICs), which are paired with each CPU. External I/O and storage-system components are implemented in custom-designed FPGAs. The FPGAs implement both the SeaMicro fabric technology and standards-based I/O technology. The fabric side of the FPGAs connects to the ASICs to form the fabric, while the standards-based I/O side connects to external interfaces providing standards-based interfaces for the network and storage cards.

To summarize, hardware-based CPU I/O virtualization enables SeaMicro drive down the number of components in a server and to shrink the motherboard to the size of a credit card. Hundreds of these low-power, card-sized computational units are tied together with a supercomputer-style fabric to create a massive array of independent but linked computational units. Work is then distributed over these hundreds of CPUs via hardware- and software-based load-balancing technology that dynamically directs load to ensure that each of the CPUs is either in its most efficient zone of performance or is sleeping. The key technologies reside in three chips of SeaMicro's design, one ASIC and two FPGAs, and in the management, routing, and load-balancing software that directs traffic across the fabric.

This combination of technologies produces current systems featuring SeaMicro technology that typically use one quarter the power and take one sixth the space of traditional servers with the same compute performance, yet deliver up to 12 times the bandwidth per core.[6]

1.7 Recommended Reading

A copy of the Final Report to the US Department of Energy entitled *Recovery Act: SeaMicro Volume Server Power Reduction Research and Development, March 22, 2012, Award Number DE-EE0002889,* can be downloaded at http:www.osti.gov/bridge

Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431, U.S. Environmental Protection Agency ENERGY STAR Program, August 2, 2007.

[6] ¼ the power based on a comparison of same throughput achieved by 51 traditional 2P 1U rack servers @ 13,581 total watts at 100% utilization and 512 1P SeaMicro servers in a single 10U chassis at 2,499 total watts at 100% utilization, running a standard web server workload. 1/6 the floor space is based on a comparison of 60 1U traditional servers versus one SeaMicro chassis at 10U. 12X throughput is based on a traditional Dual socket platform with 12 cores (2 socket x six cores) and 2x1GB NICs (2 Gig/12 cores = 167 Mbps bandwidth per core) compared to a SeaMicro single socket server 4 cores and aggregated bandwidth of up to 8 1 Gig NICs for each socket (8 Gig/4 cores = 2Gbps bandwidth per core) 2/.167 = 12

CHAPTER 2:
Redundancy and Reliability Overview

2.1 Hardware Reliability

The SeaMicro SM10000 is a 10 RU standard 19" rack mount server comprised of several major hardware blocks:

- 64 hot-swappable compute cards, each of which holds four dual core 1.66 GHz Intel Atom processors for a total of 512 cores. Two SeaMicro ASICs connect to each Dual Core CPU.
- Up to 8 hot-swappable uplink I/O cards, each of which holds either eight 1 Gigabit Ethernet uplinks or two 10 Gigabit Ethernet uplinks, for a maximum of 64 x 1 GbE ports or 16 x 10 GbE ports
- Up to 8 hot-swappable storage cards, each of which holds 8 hot swappable SATA HDD or SSDs.

The system can be configured to run with no disk, or with up to 64 disks.

These hardware blocks are tied together with a super computer-style fabric created by interconnecting 512 SeaMicro ASICs. The fabric is a kary n-cube, more commonly known as a torus, with k=8 (nodes per dimension) and n=3 (dimensions) for a total of 512 end ASICs (each CPU is paired with two SeaMicro custom ASICs and DRAM to create a "compute node"). The fabric is low latency, high bandwidth and massively fault tolerant, and looks like a 3-dimensional cubeshaped donut. The Intel Atom CPUs, Ethernet I/O, and storage cards all connect to this resilient fabric.

Packets are routed over the fabric using a strict deterministic dimensional order routing (traverse completely in one-dimension before moving to another dimension) in order to prevent fabric dead-locks. The SeaMicro fabric can be programmed for traffic to traverse a secondary path in case of primary path failure. Secondary routes can use any of the six available links from each ASIC, providing sufficient redundancy within the fabric. The SeaMicro fabric also implements packet classes and virtual channels in order to better utilize the link bandwidth and eliminate packet loops. Implementation of virtual channels and secondary routes enables a higher level of resiliency and also eliminates the need for error-prone link-level loop prevention and redundancy protocols such as the spanning tree.

2.2 I/O and Storage Cards

Connecting to the fabric at multiple locations are the I/O and storage cards. The I/O and storage cards have a similar architecture. Both the I/O and the storage card contain two FPGAs and each FPGA connect to four different "ASICs" in the fabric. In the case of a FPGA link failure, the SeaMicro fabric routing automatically forwards traffic via one of the remaining three unaffected paths.

The SM10000 can be configured with up-to 8 hot-swappable I/O cards. To protect application traffic, the Ethernet interfaces can be bundled together in a Link Aggregation Group (LAG). A LAG can be created by using interfaces on different Ethernet cards, enabling automatic link level resiliency in the case of I/O card failure. Multiple LAGs can be configured to be in a primary/secondary mode to provide for resiliency in case of a downstream switch/router failure. SeaMicro software automatically switches traffic from the primary to the secondary LAG upon detecting any failure in the primary LAG.

The SM10000 can also be configured with up to 8 hot swappable storage cards. Each storage card holds up to 8 physical disks and each physical disk can be sliced into multiple virtual disks. Individual disks or the entire storage card can be hot-swapped. The two FPGAs that reside on the storage card create the hardware path for data access from the virtual disks and implement transparent hardware RAID to protect against disk failure. In the case of a disk failure in a configured RAID group, the SeaMicro system software provides the ability to rebuild data on a newly inserted disk. In addition to RAID, multiple virtual disks can be assigned to each of the CPUs providing the ability to store data on distinct disks on different storage modules to protect against failure.

2.3 Power Supplies and Fans

The SM10000 follows the data center standard of front-to-rear air flow using two fan trays. Each fan tray hosts three dual-stacked fans to pull air from the front and exhaust through the rear. Fan trays are hot swappable and can be replaced on failure. Fan redundancy ensures that if a fan fails, the system continues to operate without impact to the system.

On the back side of the SM10000 are four independent slots for the hot-swappable AC-DC power supplies. A fully-configured system requires no more than three AC-DC power supplies providing a 3+1 redundancy. All the power supplies connect to a common passive power bus that distributes power within the system. The failure of a single power supply has no impact to system availability and operation.

2.4 Software Resiliency

In addition to hardware redundancy ensuring no single point of failure, SeaMicro provides an array of redundant and resilient software features. The SeaMicro management software runs on a dedicated CPU on the I/O card (not one of the Atom processors). Adding a second I/O card automatically provides control plane redundancy – active/standby.

In the SM10000, individual CPUs can be placed into pools of compute by the built-in hardware based load balancer. The stateful load balancer, in combination with the management software, keeps track of the health of the system, the CPUs, and the applications. The load balancer distributes work to multiple processors via userconfigured algorithms such as round-robin, least connection, and max-connection. If one of the CPUs fails, the application healthcheck automatically isolates the failed processor and notifies the load balancer to redirect user flows to other healthy CPUs in the system. In fact, SeaMicro customers can proactively isolate a CPU from user traffic in order to perform software upgrades.

The SeaMicro management software also provides users with the ability to identify processors that are running critical applications, and provision a group of “spare” CPUs. If the CPU running a critical application experiences a failure, the management software in the SM10000 automatically assigns the virtual disks and the environmental parameters to a spare processor and restores application access to the end user.

The SM10000 also protects against the possibility that the management software or the hardware that runs the management software fails. The management software is built with two levels of resiliency. First, because of its modular nature, if a software module fails it is automatically restarted without any disruption. Second, the SM10000 fully supports redundant management cards. The management software continually monitors the health of the hardware and software on the standby management card. If the hardware or software on the primary card fails, the secondary card takes over and a new secondary card is elected with no manual intervention. In the case of either failure, SeaMicro software sends out failure notification through the management user interfaces. The system also maintains sufficient error logs for debugging and troubleshooting purposes.

CHAPTER 3:
Test Plan to Show that SeaMicro System Saves Power and Space

This plan verifies that current systems featuring SeaMicro technology typically use one quarter the power and take one sixth the space of traditional servers with the same compute performance.

3.1 Test Plan

Following is the plan to measure the overall throughput of a SeaMicro Chassis and compare that with commodity hardware. For this case, we will be using a standardly equipped Dell R410.

Both hosts were set up with Apache 2.2 and saturated with http requests to derive an accurate throughput value for each. During this run we will have a power meter hooked up to help measure the overall power usage of the chassis.

Figure 1: Plan to Measure throughput of SeaMicro Chassis

3.2 Methodology

To accomplish this comparison SeaMicro setup a standard web workload identified as a static 16KB file. We then used Apache Bench to retrieve this file over a set period of time to get an accurate maximum throughput measurement for each host. This data was then extrapolated out to show the energy saving a SeaMicro chassis can provide.

SeaMicro used two different methods to showcase this data. The first method to provide a linear scaling model that assumes that for every network interface card the user adds it will obtain twice the bandwidth. The second method to show the data in a more real world scenario where for every network card the user adds it will obtain a 60 percent return on overall throughput. SeaMicro chose these two data points because from prior tests it was shown that the onboard network interface cards included in today's commodity hardware do not scale linearly.

3.3 Problem and Solution Statement

Current commodity hardware has lots of redundancy because every server is separate and thus requires lots of duplicate hardware such as power supply, fans, and motherboard components. The solution to this problem is to provide a system, which eliminates all this extra hardware and uses a lower power CPU to further the benefit.

Table 1: Configuration & Technical Specifications

Configuration	Technical Specifications
Dell R410: - *2 x L5630 2.13 GHz Quad core Xeon* - *24GB of ram* - *Single Power supply 480watts* - *Dual onboard Gigabit NC* SeaMicro SM1000: - *512 Z530 1.6 GHz Atom* - *2GB of ram per CPU* - *Redundant power supply* - *64 Gigabit ports* - *1.28 Tb/s internal Fabric*	*Apache* 2.2.3 *OS* CentOS 5.4 *Payload Size* 16KB *Test length* 600sec *kWs used by one SeaMicro host* 2.5kW *Kilowatts used by Dell hosts* 18.4kW *Dell throughput per host* 9791rps *SeaMicro throughput per host* 1963rps

Table 2: Results

Measured Data	Form Factor	Hosts	Throughput/sec	Power	CPU util
Dell R410 1Gig Nic	1U	1	5,699	168	20%
Dell R410 2Gig Nic	1U	1	9,791	180	30%
SeaMicro Single node	10U	1	1,963	4.88	100%
Normalized Linear Scaling 100% CPU Util.	**Rack Space used**	**Hosts**	**Throughput/sec**	**Power**	**SeaMicro Power Benefit**
Dell R410	34U	31	1,005,056	8,449	70%
SeaMicro	10U	512	1,005,056	2,499	
SeaMicro Space Benefit	71%				
Normalized Non-linear Scaling 100% CPU Util.	**Rack Space used**	**Hosts**	**Throughput/sec**	**Power**	**SeaMicro Power Benefit**
Dell R410	54U	51	1,005,056	13,581	82%
SeaMicro	10U	512	1,005,056	2,499	
SeaMicro Space Benefit	81%				

*Note: A total of 750 watts and 3U were added to each Dell scenario for two extra switches and terminal server hardware. For non-linear scaling we used 58% as documented in the "Measured Data" table. This data was measured with a single power supply in the Dell. If we changed the configuration to have a redundant power supply the power benefit would be even greater.

3.3.1 Analysis and Conclusion

Current systems featuring SeaMicro technology typically use one quarter the power and take one sixth the space of traditional servers with the same compute performance, yet deliver up to 12 times the bandwidth per core.[7]

[7] ¼ the power based on a comparison of same throughput achieved by 51 traditional 2P 1U rack servers @ 13,581 total watts at 100% utilization and 512 1P SeaMicro servers in a single 10U chassis at 2,499 total watts at 100% utilization, running a standard web server workload. 1/6 the floor space is based on a comparison of 60 1U traditional servers versus one SeaMicro chassis at 10U. 12X throughput is based on a traditional Dual socket platform with 12 cores (2 socket x six cores) and 2x1GB NICs (2 Gig/12 cores = 167 Mbps bandwidth per core) compared to a SeaMicro single socket server 4 cores and aggregated bandwidth of up to 8 1 Gig NICs for each socket (8 Gig/4 cores = 2Gbps bandwidth per core) 2/.167 = 12

CHAPTER 4:
Case Studies

4.1 Mozilla Slashes Power Draw and Operating Expense While Increasing Compute Density and Reliability with the SeaMicro SM10000

Mozilla is a public benefit, non-profit organization responsible for the popular Firefox browser. Mozilla is built from a global community of people who believe that openness, innovation, and opportunity are key to the continued health of the Internet. Firefox, Mozilla's flagship product, is free and open source software, with approximately 40 percent of its code written by volunteers. Today, more than 400 million people around the world use Firefox.

Justin Fitzhugh leads the team responsible for the engineering and infrastructure that supports this global user base. Mozilla's business depends on a compute infrastructure that is available to Mozilla's own engineers, to open source developers around the world, and to millions of end users each day.

"The primary challenges of managing a large compute infrastructure are power and space. They are my dominant concerns," said Fitzhugh. "Power determines how many servers you can fit in your facility, which in turn determines how many users or engineers you can support. Inside your power envelope, server density determines how many servers can fit in a given facility and how many facilities you require. At the end of the day, the power and space used by servers dominates the total cost of ownership in a data center."

Mozilla had long known that the traditional blade servers they were purchasing were ill suited for their workload but saw no alternative. Like most Internet companies, Mozilla runs Linux, PHP and Apache. Their workloads were comprised of millions of relatively small, independent jobs, like an individual Firefox end user checking for an update or adding a plug in. Traditional servers were built for a very different workload — workloads that were large, interrelated, and complex. While Mozilla recognized the mismatch between the blade servers they were purchasing and their LAMP stack workload, there were no alternatives in the market.

"Innovation in the server space has been rare. In SeaMicro, we saw something transformative. A company committed to slashing power draw and dramatically increasing compute density, while remaining standards based," said Fitzhugh. "We saw engineers attacking problems that other companies had refused to tackle. We knew they were on to something special."

SeaMicro servers are optimized for the Internet workload that dominates Mozilla's infrastructure. The SM10000 integrates 512 Intel® Atom™ low power processors, traditionally used for mobile devices and the smallest laptops. The SeaMicro architecture links these processors together with a super-compute style fabric transforming the system into a high density, low power, single-box cluster computer, optimized for Internet traffic. SeaMicro further improved total cost of ownership and reduced power use by integrating the functionality traditionally found in an entire data center rack — compute, storage, networking, server management, and load balancing – into a single, low-power system.

Mozilla performed extensive testing prior to deploying the SeaMicro SM10000. Mozilla's priorities were performance/unit power, ease of use and operation, and reliability. Mozilla tested and compared the SeaMicro system to their existing blade enclosures and discovered enormous compute per unit power advantages at all CPU utilizations.

Today, when an end user running Firefox wants to download an update, addon or persona, the browser sends a URL request to Mozilla's download cluster. The servers running Linux, Apache, and PHP run a query to the backend database/memcache cluster and return an HTTP Redirect URL (also called an HTTP 302 return) back to the browser so that Firefox can fetch the desired file from the network of donated download resources.

Mozilla compared performance on the SeaMicro SM10000 with their incumbent system - an HP C7000 Dual Socket Quad Core L5530 Xeon Blade and found SeaMicro to be dramatically superior on all of the competitive dimensions: Capital expense per unit compute, HTTP requests serviced per/Watt, and HTTP requests serviced per Rack Unit. These advantages combined to produce dramatic savings in both capital expense and operating expense. On the operating expense side, the SeaMicro SM10000 used less than 1/5 the power per request, and took less than 1/4 the space of the HP C7000 for small transaction, high volume workloads.

In addition, while handling HTTP requests, each Atom server in the SeaMicro SM10000 provided dramatically more consistent response time to user requests than the Quad Core Xeon processors in the HP C7000, ensuring a uniform and positive customer experience.

"Each time you download one of the 10,000 Firefox plug-ins, you are traversing SeaMicro equipment. The SM10000 was easy to deploy, improved compute per unit power, and simplified our infrastructure by eliminating Ethernet switch ports and reducing management overhead," said Mathhew Zeier, Director of Operations.

Today SeaMicro's SM10000 is an important part of the Mozilla infrastructure. Said Fitzhugh, "Our investment in SeaMicro has already paid tremendous dividends. We are drawing on the order of 1/5 the power per HTTP request, and using less than 1/4 the space. This has freed room in our power envelope, and space in our cages to increase the compute footprint, providing additional CPU cycles for our different users, while dramatically reducing our operating expenses."

4.2 eHarmony Success Story

4.2.1 SeaMicro's SM10000™ Servers Increases eHarmony's Compute Time and Reliability While Reducing Total Cost of Ownership

eHarmony (www.eharmony.com) is a pioneer in using relationship science to match singles seeking long-term relationships. Its service, available in the United States, Canada, Australia, the United Kingdom and Brazil, presents users with compatible matches based on key dimensions of personality that are scientifically proven to predict highly successful long-term relationships. An average of 542 eHarmony members marry daily in the United States as a result of being matched on the site*.

Each day, eHarmony sifts through data provided by its millions of members to find the best possible matches. Cormac Twomey is responsible for implementing the matching algorithms and manages the extensive compute infrastructure required to find compatible matches among eHarmony's users. The data set is large and complex, and the algorithms used are extremely sophisticated.

To achieve matching on this massive scale, the application developers at eHarmony turned to Apache Hadoop. Hadoop is an open source technology that allows many small independent servers to work together as if in a cluster to solve big complex problems. Hadoop enables applications to work with thousands of compute nodes and petabytes of data. Initially, to meet its compute requirements, eHarmony turned to a leading cloud provider and rented compute by the hour. The cloud environment provided a flexible but very expensive solution for eHarmony's Hadoop cluster.

As eHarmony's membership grew exponentially, it sought a more cost-efficient solution for its Hadoop work. After extensive testing, eHarmony purchased the SeaMicro SM10000™ high density, low power server with a configuration that enabled its Hadoop application to complete its run in exactly the same time that it had been taking in the cloud. In comparison with its previous four hours per evening cloud-based rental, eHarmony reduced its operating expenses by tens of thousands of dollars a month, and its TCO by more than 74 percent.

"Each month, we were paying huge amounts for a few hours per night of compute in the cloud," said Twomey. "We brought the SeaMicro system in-house, and this allowed us to save tens of thousands of dollars each month, complete our Hadoop run in the same elapsed time, and use the rest of the day of compute for free."

In addition to the TCO savings, eHarmony benefited from the following SeaMicro capabilities:

- Additional compute: For less than they were paying for a few hours of cloud compute, eHarmony was able to purchase the SeaMicro SM10000 and have it running 7 x 24 in its facility. This gave eHarmony the opportunity to run its Hadoop jobs more often without any additional costs.

- Reliable completion time: In the cloud, the amount of time it took Hadoop to complete the run varied widely, making it difficult to predict when the matching data could be delivered to other parts of the organization. This variability stems from the fact that in the cloud, CPUs, DRAM, disk, and network bandwidth are shared among customers competing for resources. On the other hand, inside the SeaMicro SM10000, resources were dedicated to eHarmony, and the system was able to reliably replicate run times. This allowed the operations team to keep its commitments to other parts of the organization and even offer SLA's to colleagues.

- Eliminated data upload charges: Cloud providers charge customers a fee to upload data. Loading data into the SeaMicro system was quick and easy and without charge.

- Accelerated application performance: eHarmony was able to leverage SeaMicro’s 512 compute nodes and its high performance fabric to reduce data retrieval latency between its Hadoop and memcached applications by running both in the same SeaMicro system.

- Simple management: With 512 Intel® Atom™ processors inside the SeaMicro SM10000, the management, cabling, and switching hassles of building a cluster were removed. The SM10000 provided a single box Hadoop cluster that was easy to install and simple to manage.

Twomey concluded, “In the end, it seemed like a no-brainer. We purchased SeaMicro servers and immediately reduced our operating expenses, delivered our Hadoop work more reliably, impressed our colleagues with an internal SLA, and had more compute time available to refine and improve our methodologies. I wish more decisions were this easy and paid dividends this quickly.”

CHAPTER 5:
SeaMicro's Updated Compute Card to Address New Markets

The SM10000-64HD combines SeaMicro's revolutionary server architecture with 384 of the latest 64-bit Intel® Atom™ N570 dual-core processors. The resulting system delivers an industry best 768 1.66 GHz x86 cores in 10 rack units (17.5 inches). The previous industry record was held by SeaMicro's SM10000-64, which sports 256 dual-core Atom processors for a total of 512 1.66 GHz x86 cores in 10 rack units. Bringing together Intel's most advanced Atom processor and SeaMicro's Internet-optimized server architecture, the SM10000-64HD combines the benefits of energy efficiency, high compute density, and 64-bit x86 software support.

Like all SeaMicro servers, the SM10000-64HD is exceptionally easy to use. It simplifies data center operations and management by eliminating layers of Ethernet switches, server management devices and expensive load-balancers. It is plug and play – customers can deploy it without modifications to existing operating systems, application software or management tools.

SeaMicro's SM10000-64HD system is comprised of:

- 384 x86 dual-core 1.66 GHz Intel Atom processors,
- 1.536 terabytes of DDR3 DRAM,
- up to 64 SATA solid state or hard disk drives,
- 8 - 64 one gigabit Ethernet uplinks.

Additional benefits include:

- support for 64-bit operating systems,
- four gigabytes of addressable memory per socket,
- industry leading density with 768 cores per 10 rack unit, or 3,072 cores per 7-foot rack,
- runs off-the-shelf OSs for application compatibility, enabling drop-in adoption,
- integrated top-of-rack switch, load balancer and console server, simplify infrastructure by eliminating equipment that adds unnecessary cost and management complexity.

APPENDIX A:
SeaMicro Photos and Visual Aid

Figure A1: SeaMicro Building Headquarters

Figure A2: Original Computer Card

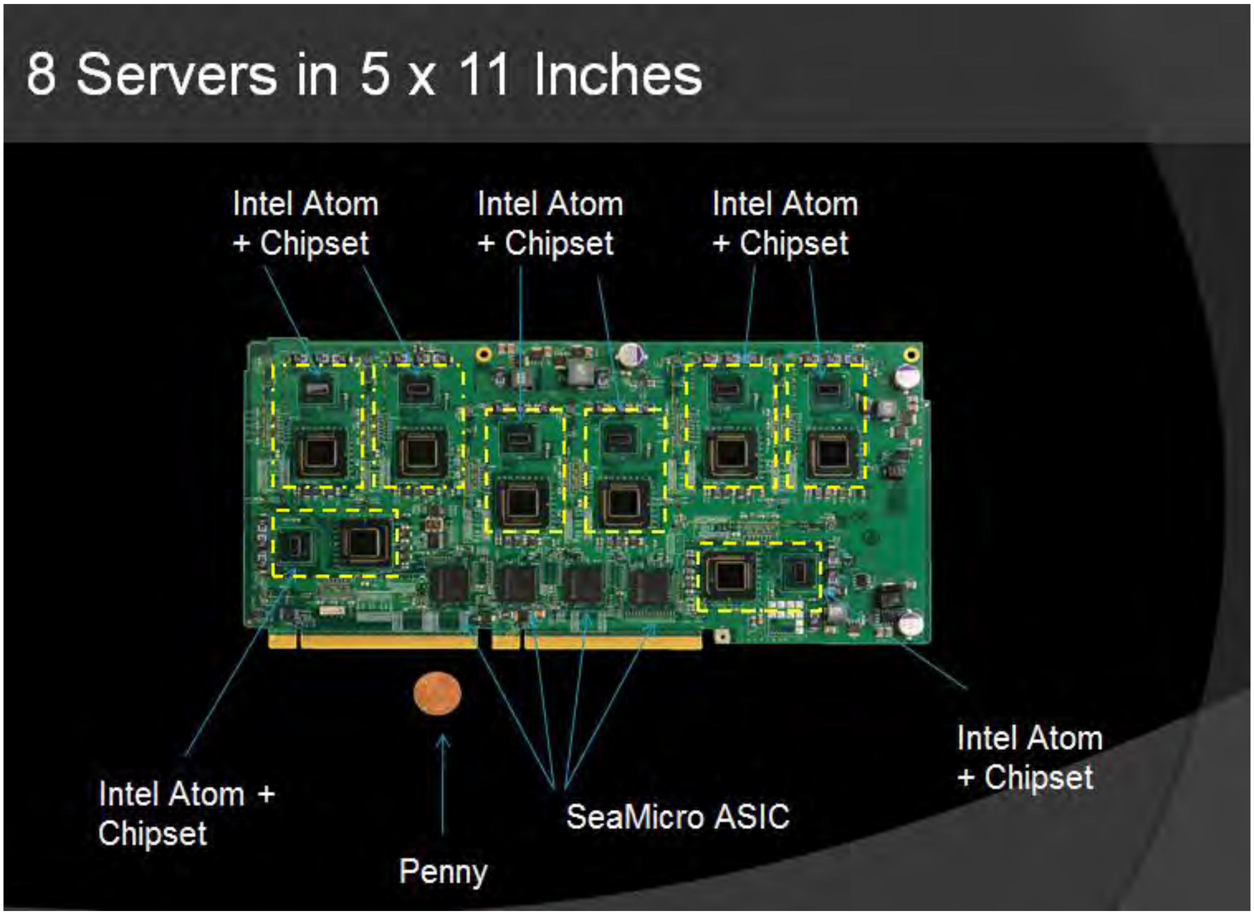

Figure A3: Six Dual Core Servers in 5x11 inch Space

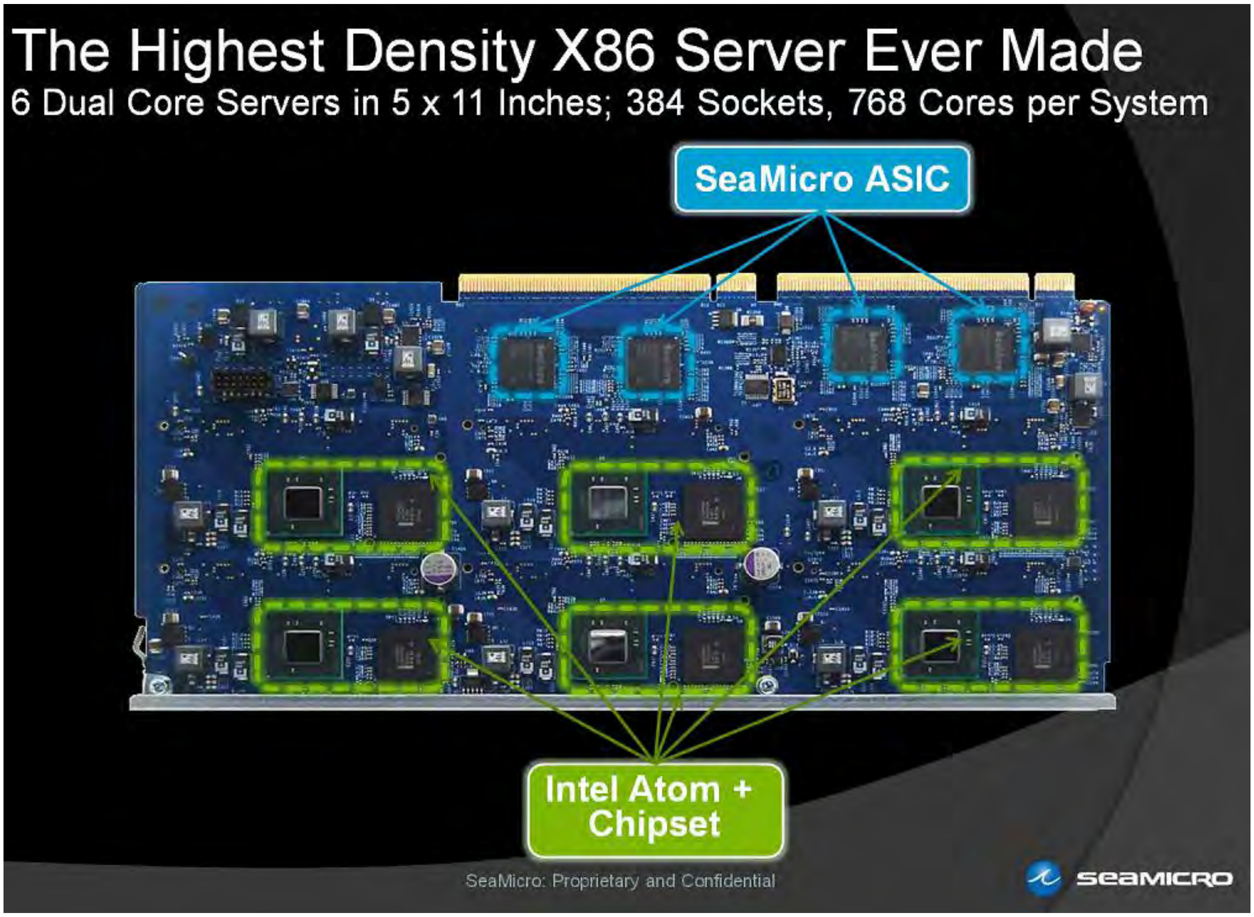

Figure A4: SeaMicro Compute Appliance

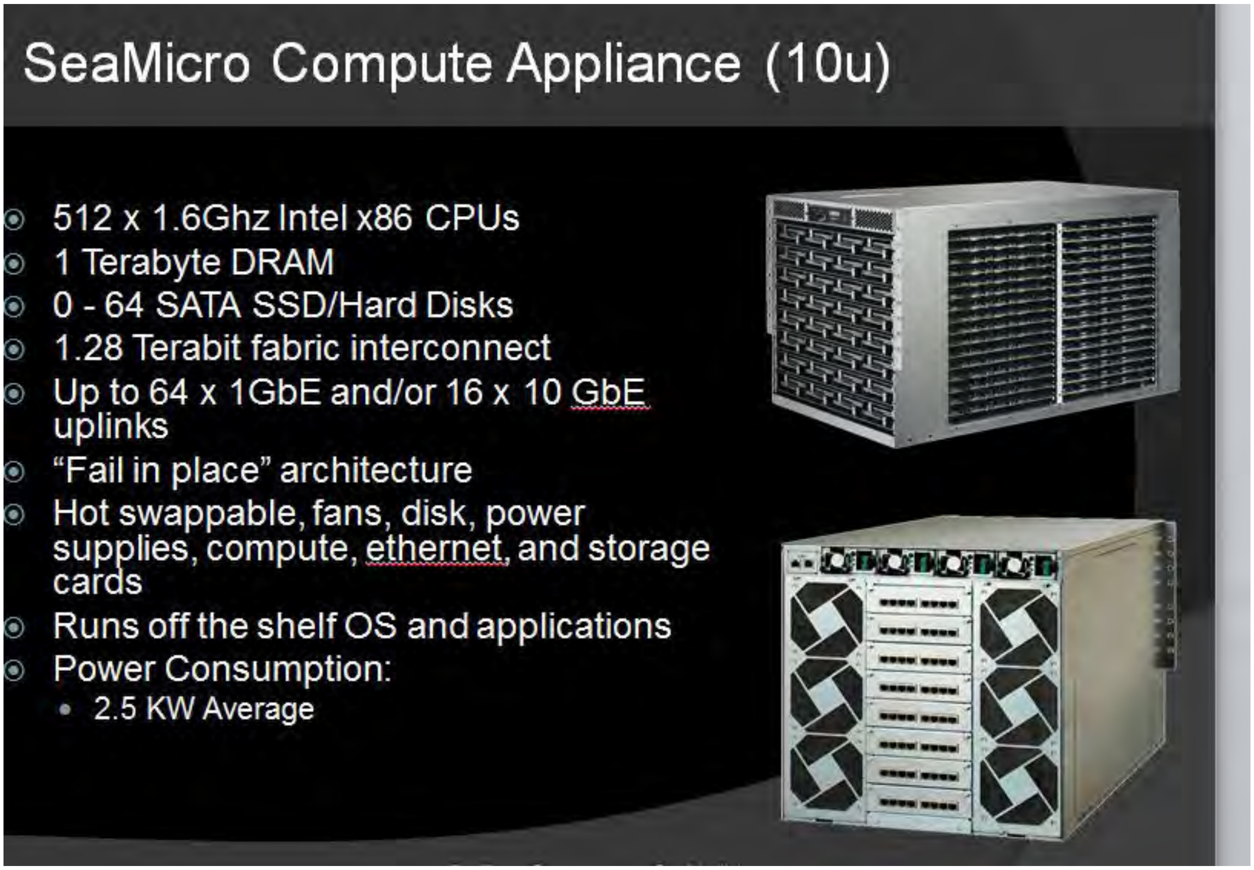

Figure A5: Total Cost of Ownership